Individual Assignment 7

Please download the **Star Wars.csv** dataset from Canvas.

This file contains movie goer survey data. Disney has commissioned a survey of moviegoers regarding their attitudes toward Star Wars properties and other movies.

Below are the column definitions for this sheet:

- Saw Most Recent: 1 if the moviegoer saw the most recent Star Wars movie, 0 if not *(output variable)*
- Fan: 1 if the moviegoer considers themselves to be a Star Wars fan, 0 if not
- Films Seen: # of Star Wars films seen prior to the most recent
- Prev Avg Rating: average rating of Star Wars movies prior to the most recent, out of a maximum score of 6
- Han Solo: sentiment towards the character Han Solo, 1 being the least positive, 3 being the most positive
- Princess Leia: sentiment towards the character Princess Leia, 1 being the least positive, 3 being the most positive
- Anakin Skywalker: sentiment towards the character Anakin Skywalker, 1 being the least positive, 3 being the most positive
- Darth Vader: sentiment towards the character Darth Vader, 1 being the least positive, 3 being the most positive
- Expanded Universe Fan: 1 if fan of the Star Wars Expanded Universe, 0 if not
- Star Trek Fan: 1 if a fan of Star Trek, 0 if not
- Age
- Household Income (in USD)
- EducationIndex: maximum education attained by the moviegoer, 1 being the least, 4 being the most
- EastCoastResident: 1 if moviegoer lives on the east coast, 0 if not

Disney wants to be able to predict which movie goers will see the next Star Wars movie based on their characteristics and relationship to Star Wars.

You will use this survey data to build models in R that can predict whether a moviegoer saw the most recent Star Wars movie or not.

Open RStudio and create an R Markdown file. Save the .rmd file, ensuring that you save it in a file location that you can easily find again.

Each of the following items corresponds to a code chunk, or text to record in the R Markdown file. Please make sure to label each item using the numbers provided in this document.

Make sure to run each code chunk after you write it to ensure that there are no errors. Be careful about typos!

If you have not installed the caret and ranger libraries yet, do so according to the instructions given in class before beginning this portion of the assignment.

Item 1, Loading Packages [Code Chunk]: Load the caret and ranger libraries.

Individual Assignment 7

Item 2, Importing Data [Code Chunk]: Import the Star Wars.csv file into an R data frame called "starwars".

Item 3, Pre-processing data [Code Chunk]: Change the Saw Most Recent variable to a categorical variable for the purpose of binary classification.

Item 4a, Partitioning Data [Code Chunk]: Set the random seed to 1234.

Item 4b, Partitioning Data [Code Chunk]: Partition training & validation datasets; put 70% of the data into training and the rest into validation.

Item 5, Normalizing Data Partitions [Code Chunk]: Normalize the training and validation data partitions. Make sure the normalized partitions are saved into new data frames as was shown in class.

Item 6a, Training k-nearest neighbors model [Code Chunk]: Train a k-nearest neighbors model on the normalized training data. Make sure to display the performance of various values of k, as well as plot the accuracy of various values of k.

Item 6b, Training k-nearest neighbors model [Text]: What value of k produces the best accuracy?

Item 7, kNN Predictions [Code Chunk]: Generate kNN predictions for the normalized validation data.

Item 8, Training random forest model [Code Chunk]: Train a random forest model on the training data partition (not normalized data)

Item 9, Random forest predictions [Code Chunk]: Generate random forest predictions on validation data partition (not normalized)

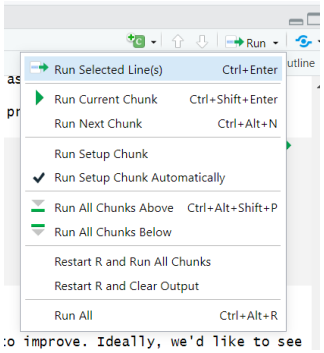
Item 10a, Validation Confusion Matrices [Code Chunk]: Produce the confusion matrix for the random forest model's predictions on the validation data.

Item 10b, Validation Confusion Matrices [Code Chunk]: Produce the confusion matrix for the k-nearest neighbor's model's predictions on the validation data.

Item 10c, Validation Confusion Matrices [Text]: Compare and contrast the performance of the k-nearest neighbors and the random forest models. Which model would you use? Use at least 2 performance metrics to support your answer.

Click "Run All" to run all of your code chunks and check the output. Make sure that any text answers you have match this output. This is to ensure that your text answers match the output produced by R, factoring in randomization!

Individual Assignment 7



Save your file, then knit and export your R Markdown file as an HTML file. Upload the HTML file to Canvas to complete the assignment.

If you are unable to knit your .rmd file due to errors, make sure that you go back and test your code chunks individually. If you are ultimately unable to figure out how to solve these errors, save the .rmd file and upload that instead of the HTML file for partial credit.

If you are unable to find your exported HTML file, consult Canvas for instructions.